Problem 1A

1.1) I) One way ANOVA hypothesis for Education

Null hypothesis

H0: X1=X2=X3

Alternate hypothesis

HA: X1!=X2!=X3

OR

Mean of any one pair is unequal

Here, X1= Mean salary of High school graduates

X2= Mean salary of Bachelors

X3= Mean salary of Doctorates

II) One way ANOVA hypothesis for Occupation

Null hypothesis

H0: Y1=Y2=Y3=Y4

Alternate hypothesis

HA: Y1!=Y2!=Y3!=Y4

OR

Mean of any one pair is unequal

Here, Y1= Mean salary in Administrative and clerical

Y2= Mean salary in Sales

Y3= Mean salary in Professional or speciality

Y4= Mean salary in Executive or managerial

1.2) P value = 1.257709e-08

Level of significance = 0.05

As P value is smaller than level of significance, the Null hypothesis is rejected.

1.3) P value = 0.458508

Level of significance = 0.05

As P value is greater than level of significance the null hypothesis is accepted.

Problem 1B

1.5) Hypothesis of interaction between education and occupation on salary

Null hypothesis : There is no interaction between education and occupation

Alternate hypothesis :There is interaction between education and occupation

P value = 2.232500e-05

Level of significance = 0.05

As P value is lesser than level of significance null hypothesis is rejected i.e there is interaction between education and occupation on salary.

The point plot between education and occupation shows the following interaction:

i) With Doctorate degree salary is maximum in Prof-speciality and minimum in Adm-clerical department.

ii) With Bachelors education salary in Exec-managerial and Sales department are equal and maximum and minimum in Prof-speciality department..

iii) With HS-grad salary is maximum in Prof-speciality and minimum in Sales department.

iv) In the Prof-speciality department there is a lot decrease in salary from a Doctorate to Bachelors , but decrease in salary between Bachelors education and HS-grad is not much.

v) In the Sales department salary of Doctorate and Bachelors are equal but for HS-grad it is less.

vi) In the Adm-clerical department salary of Bachelors is little more than doctorate and HS-grad is least.

vii) in the Exec-managerial department salary of Doctorates is slightly higher than Bachelors.

1.6) Null hypothesis : mean salary at all education and all occupation level are same

Alternate hypothesis : mean salary at all education and occupation level are not equal

After performing 2 way Anova:

i) P value of Education = 1.981539e-08

P value of occupation = 3.545825e-01

Level of significance=0.05

As P value of both Education and occupation are less than significance level , Null hypothesis is rejected which implies that mean salary depends on education and occupation level.

Hypothesis of interaction between education and occupation on salary

Null hypothesis : There is no interaction between education and occupation

Alternate hypothesis :There is interaction between education and occupation

P value = 2.232500e-05

Level of significance = 0.05

As P value is lesser than level of significance null hypothesis is rejected i.e there is interaction between education and occupation on salary.

1.7)i) Doctorates are costing more to the company and HS-grads are costing less to the company.

ii) in departments executive-managers are more costly.

Problem 2

2.1) The data has been downloaded correctly.

There are 777 rows ranging from 0 to 776 and 18 columns.

There are no null values,missing values and duplicates in the data set.

Data type of 1 column(Name) is object, 1 column (S.F.Ratio) is float and remaining columns are integer

i)Univariate analysis

- The mean of total applications received is 3001.638353

Median is 1558

Minimum is 81

Maximum is 48094

The data is not normally distributed ,is positive skewed and contains outliers on the higher side.

-The mean total number of applications accepted is 2018.80437

Median is 1110

Minimum is 72

Maximum is 26330

The data is not normally distributed, is positive skewed and contains outliers on the higher side.

-The mean of total new students enrolled is 779.972973

Median is 434

Minimum is 35

Maximum is 6392

The data is not normally distributed , is positive skewed and contains outliers on the higher side.

-The mean percentage of new students from top 10% of higher secondary class is 27.558559

Median is 23

Minimum is 1

Maximum is 96

The data is almost normally distributed with slightly positive skewness and contains outliers on the higher side.

-The mean percentage of new students from top 25% of higher secondary class is55.796654

Median is 69

Minimum is 9

Maximum is 100

The data normally distributed and contains no outliers.

-The mean number of full -time undergraduate students is 3699.907336

Median is 1707

Minimum is 139

Maximum is 31643

The data is not normally distributed ,is positive skewed and contain outliers on the higher side

-The mean number of part-time undergraduate students is 855.298584

Median is 353

Minimum is 1

Maximum is 21836

The data is not normally distributed ,is positive skewed and contains outliers on the higher side.

-The mean number of students for whom the university is out of station tuition is 10440.669241

Median is 9990

Minimum is 2340

Maximum is 21700

The data is normally distributed and contains only one outlier on the higher side.

-The mean cost of room and board is 4357.526384

Median is 4200

Minimum is 1780

Maximum is 8124

The data is almost normally distributed with slight positive skewness and contains outliers on

the higher side.

-The mean book cost for students is 549.380952

Median is 500

Minimum is 96

Maximum is 2340

The data is not normally distributed , is positive skewed and contains outliers on both higher and lower sides.

-The mean personal expenditure per student is 1340.642214

Median is 1200

Minimum is 250

Maximum is 6800

The data is not normally distributed , is positive skewed and contains outliers on

the higher side.

-The mean percentage of faculties with phd is 72.660232

Median is 75

Minimum is 8

Maximum is 103

The data is not normally distributed , is negative skewed and contains outliers on the lower sides.

-The mean percentage of faculties with terminal degree is 79.702703

Median is 82

Minimum is 24

Maximum is 100

The data is not normally distributed , is negative skewed and contains outliers on the lower sides.

-The mean student/faculty ratio is 14.089704

Median is 13.6

Minimum is 2.5

Maximum is 39.8

The data is not normally distributed , is slightly positive skewed and contains outliers on both upper and lower sides.

-The mean percentage of alumni who donates is 22.743887

Median is 21

Minimum is 0

Maximum is 64

The data is almost normally distributed with slightly positive skewness and contains outliers on the upper side.

- The mean instructional expenditure per students is 9660.171171

Median is 8377

Minimum is 3186

Maximum is 56233

The data is not normally distributed , is positive skewed and contains outliers on the upper side.

-The mean graduation rate is 65.46332

Median is 65

Minimum is 10

Maximum is 118

The data is almost normally distributed with slight negative skewness and contains outliers on both lower and upper side.

2.2) As values of all variables are not of the same weight i.e some values are in 2 digits and some values are in 4 or 5 digits , this data requires scaling before performing PCA. Scaling has been done by using z score.

2.3) covariance and correlation matrices are showing following relations:

* Application applied and application accepted are showing positive, linear relation.As application applied increases acceptance of application also increases.
* Applications accepted and enrollment are showing positive, linear relation. As acceptance of applications increases enrollment also increases.
* Enrollment and full time graduates are showing positive and linear relation. Most of the enrollment have opted for full time.
* Phd and terminal are showing linear and positive relation.
* Full time graduates are showing positive and linear relation between applications applied,accepted and enrolled.

2.4) Insight- on scaling of data means of all variables have become equal to 0 but it has not treated outliers. Outliers are present in datasets both before and after scaling.

2.5) PCA is performed by using sklearn and PC scores are transferred into a data frame in the jupyter notebook.

2.6) Extraction of eigenvalues and eigenvectors are done in the Jupyter notebook.

2.7) 1st PC= -0.25\*SApps + 0.33\*SAccept - 0.06\*SEnroll + 0.28\*STop10perc - 0.01\*STop25perc -0.02\*SF.undergrad - 0.04\*SP.undergrad + 0.1\*SOutstate + 0.09\*SRoom.Board - 0.05\*SBooks + 0.36\*SPersonal - 0.46\*SPhd + 0.04\*STerminal - 0.13\*SS.F.Ratio + 0.08\*SPerc.alumni - 0.6\*SExpend + 0.02\*SGrad.Rate

Here S is for scaled variables.

2.8)- 1st 7 principle components can explain 85% of total variation. Hence it is sufficient to use only 7 principal components instead of 17 variables to explain the data.

-Eigen vectors indicate the direction of the principal components.we multiply eigenvectors with original data to reorient our data onto new axes.

2.9) PCA will help to reduce the dimension of data from 17 columns to lesser columns.Here we can use at least 7 PC to explain maximum variation in data. So by looking at fewer variables we will be able to know the variations in data. In this case by looking at fewer variables we will be able to know about admissions in universities, about expenditures of students and about faculties .

PCs will help to determine relations between variables and PC.